THE EFFECT OF TARGET-BACKGROUND SYNCHRONICITY ON SPEECH-IN-SPEECH RECOGNITION

Susanne Brouwer[1] *& Ann R. Bradlow*[2]

[1]Utrecht University, [2]Northwestern University
s.m.brouwer@uu.nl

ABSTRACT

The aim of the present study was to investigate whether speech-in-speech recognition is affected by variation in the target-background timing relationship. Specifically, we examined whether within trial synchronous or asynchronous onset and offset of the target and background speech influenced speech-in-speech recognition. Native English listeners were presented with English target sentences in the presence of English or Dutch background speech. Importantly, only the short-term temporal context –in terms of onset and offset synchrony or asynchrony of the target and background speech– varied across conditions. Participants' task was to repeat back the English target sentences. The results showed an effect of synchronicity for English-in-English but not for English-in-Dutch recognition, indicating that familiarity with the English background lead in the asynchronous English-in-English condition might have attracted attention towards the English background. Overall, this study demonstrated that speech-in-speech recognition is sensitive to the target-background timing relationship, revealing an important role for variation in the local context of the target-background relationship as it extends beyond the limits of the time-frame of the to-be-recognized target sentence.

Keywords: Speech-in-speech recognition, informational masking

1. INTRODUCTION

Listeners recognize speech most often under adverse listening situations in which speech may be highly degraded relative to optimal communicative settings. One common source of adverse condition includes the presence of background speech in the auditory environment. Background speech can interfere with speech recognition via energetic masking (i.e. involve overlap in the spectro-temporal content of the target and background speech signals) and/or informational masking (i.e. arise from competing demands on central processing resources). The current work addresses the impact of informational masking from background speech on target speech recognition. More specifically, we are interested in the target-background temporal relationship as a potential source of target-background contrast that can facilitate target from background segregation.

Previous work has shown several target-background variations that enhance or inhibit speech-from-speech segregation. Studies using multi-talker background babble that matched the language spoken in the target (e.g. English-in-English) have, for example, shown a release from masking when the target and background speech are spatially separated [e.g. 8] or when the gender of the target and background talkers differed [e.g. 7]. Other studies have directly compared speech recognition when the target and background languages matched or mismatched [e.g. 6, 9]. In general, these studies showed that listeners perform better on trials in which the target and background languages mismatched (e.g. English-in-Dutch) versus matched (English-in-English).

The aforementioned studies indicate several dimensions of target-background contrast that influence speech-in-speech recognition (e.g. spatial location, talker gender, language-being-spoken). However, they all involved manipulation of the target and/or the background speech which necessarily results in simultaneous variation of both energetic and informational aspects of the target-background relationship.

In an effort to understand the influence of target-background mismatch under conditions of controlled energetic masking, another research strategy is to compare speech-in-speech recognition for a fixed set of target-background pairs under varying contextual conditions. For example, [5] demonstrated variation in speech recognition accuracy for a fixed set of English-in-Dutch test trials depending on whether these test trials were presented in the context of surrounding

trials that either matched or mismatched the test trials. That is, the conditions involved either background language consistency ("pure" condition in which both test and surrounding trials were English-in-Dutch trials) or uncertainty ("mixed" condition in which test trials were English-in-Dutch trials but surrounding trials were English-in-English trials). Recognition accuracy of the English-in-Dutch test trials decreased when the test trials were presented in the mixed condition compared to the pure condition, demonstrating an influence of variation in across-trial context on speech-in-speech recognition.

Thus, this study suggested that listeners' attention to the background is quite difficult to suppress, and that variation on a relatively broad time-scale (i.e. beyond the time frame of an individual speech-in-speech test trial) is an important dimension of target-background variation for speech-in-speech recognition accuracy. Background speech variation at this broad, across-trial time-scale seems to capture attention to the detriment of target speech recognition even with controlled energetic masking (i.e. consistent target-background pairings).

In the present study, we aimed to extend our understanding of contextual influences on speech-in-speech recognition at a narrower time-scale. Specifically, we asked whether within-trial synchronous or asynchronous onset of the target speech and the background speech influenced speech-in-speech recognition. One possibility is that asynchronous target and background speech onsets and offsets may allow listeners to build up a separate stream for the background and target signals, thereby allowing listeners to more effectively tune into the target and tune out the background (i.e. listeners will show better target speech recognition accuracy under asynchronous than synchronous onset conditions). Alternatively, asynchronous onsets and offset of the target and the background speech might draw and retain listeners' attention to the background speech instead of to the target speech, and therefore listeners may show worse target speech recognition accuracy under asynchronous than synchronous onset conditions.

We tested these two possibilities for both English-in-English and English-in-Dutch sentence recognition. We included both matched (English-in-English) and mismatched (English-in-Dutch) conditions so that the magnitude of any observed influence of synchronicity could be compared to the expected replication of the target-background language (mis)match effect. Importantly, in all conditions of the present study, the energetic masking of the background on the masker remained constant; only the short-term temporal context, in terms of onset and offset synchrony or asynchrony of the target and background speech, varied across conditions.

2. EXPERIMENT

2.1. Method

2.1.1. Participants

Sixty-four native English listeners (39 females, age range 18 to 26 years) were tested. They reported not having any hearing or speech impairments. Sixteen listeners participated in each of four conditions, making this a between-subjects design.

2.1.2. Material

Three native American-English talkers and two native Dutch talkers produced the target and background stimuli. One of the English talkers provided the target speech, while the other two English talkers provided the background English speech. The two Dutch talkers provided the background Dutch speech.

Eight lists of English target sentences were selected from the revised Bamford-Kowal-Bench Standard Sentence Test [4] as target sentences. Each list contains 16 sentences with 3 or 4 keywords for a total of 50 keywords per list.

For the English background babble, we selected 200 English meaningful sentences from the Harvard/IEEE sentence lists [10]. All sentences were translated into Dutch for the Dutch background babble. From each of the 4 background talkers' recordings (2 English and 2 Dutch talkers), 100 of the 200 sentences were pseudorandomly selected, resulting in 4 different 1-talker tracks (2 in English and 2 in Dutch). Two-talker background babble tracks were then created by mixing the talkers of the same language into one single audio file in Audacity©. Both tracks were equalized to the same rms level and the long term average speech spectra of the two tracks were normalized as a means of reducing unequal amounts of energetic masking between the English-in-English and English-in-Dutch conditions. The play-out level of the target

sentences was fixed at 65 dB SPL. The background 2-talker babble tracks were played out at 68 dB SPL to produce SNRs of -3 dB when mixed with the target sentences.

The target sentences were mixed online with the background tracks using Max/MSP©. On each trial, a random portion of the desired background track was selected. In the asynchronous condition, the background babble came on 500 ms before the target sentence and continued for 500 ms after it. In the synchronous condition, the background tracks initiated and ended at the same time as the target signal (see Fig. 1). The combined target and background tracks were played out diotically over headphones.

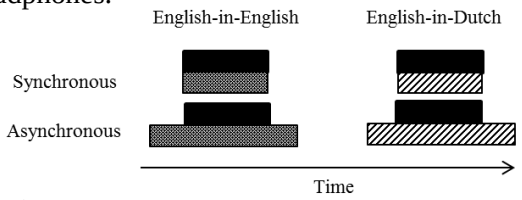

Figure 1: Illustration of the target-background timing per condition (synchronous vs. asynchronous, English-in-English vs. English-in-Dutch).

2.1.3. Procedure, design, and analysis

Listeners were instructed to listen to English sentences spoken by a native English speaker in the presence of background speech (2-talker babble). They were asked to repeat what they heard orally. A practice session of 8 trials familiarized the participants with the target talker. They were instructed to focus on the English voice that sounded less loud. Within these 8 trials all participants were able to repeat back the correct target voice. The test session included a total of 128 experimental items.

The English target sentences were presented in four separate conditions (16 listeners per condition): (1) synchronous, English-in-English, (2) asynchronous, English-in-English, (3) synchronous, English-in-Dutch, (4) asynchronous, English-in-Dutch. Thus, the background language and the synchronous and asynchronous trials were blocked. Each test session took about 25 minutes.

Data were analyzed using a linear mixed-effects regression model [1, 2] with keyword identification as the dichotomous dependent variable. A logistic linking function was used to deal with the categorical nature of the dependent variable. We constructed a 2x2 model with Background Language as one contrast-coded effect (Dutch vs. English) and Synchronicity as the other (asynchronous vs. synchronous). The Background Language by Synchronicity interaction was also included. Random intercepts were included for participants and items, along with a random slope for Synchronicity by items. Significance was assessed via likelihood ratio tests comparing the full model to a model lacking only the fixed effect [3]. In this model, a main effect of Background Language would be evidence for a replication of the mismatched language benefit [6], and a main effect of Synchronicity would be evidence for an influence of asynchronous versus synchronous presentation of the target and background speech. A significant interaction between Background Language and Synchronicity would suggest that the influence of local context (target-background onset asynchrony) is modulated by the target-background relationship within the time-frame of the to-be-recognized sentence.

2.2. Results

Figure 2 shows recognition accuracy scores for both English-in-English and English-in-Dutch trials across the asynchronous and the synchronous conditions. Black bars show average performance in the English-in-English asynchronous condition (M=61%) and average performance in the English-in-Dutch asynchronous condition (M=84%). White bars show average performance in the English-in-English synchronous condition (M=76%) and average performance in the English-in-Dutch synchronous condition (M=84%).

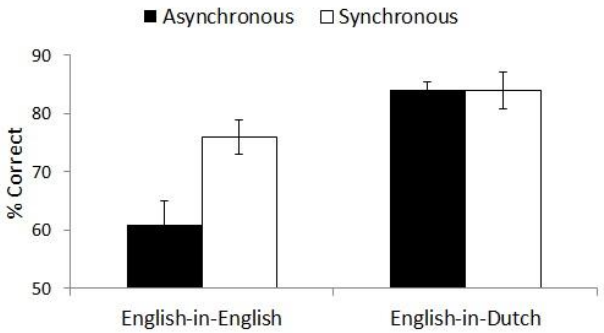

Figure 2: Mean percentage correct keyword identifications scores for the English-in-English and English-in-Dutch condition. Error bars represent standard error. (Note: data from the asynchronous conditions appeared in [5].)

The analysis showed a main effect of Background Language (β=0.15, s.e.=0.03, $\chi^2(1)$=19.93, p<0.0001), Synchronicity (β=-0.08, s.e.=0.03, $\chi^2(1)$=6.08, p<0.05) and an interaction

effect between Background Language and Synchronicity (β=0.14, s.e.=0.06, $\chi^2(1)$=5.21, p<0.05). Follow-up regressions revealed that this interaction reflected a significant effect of synchronicity for the English-in-English (β=-0.15, s.e.=0.05, $\chi^2(1)$=7.43, p<0.01), but not for the English-in-Dutch conditions ($\chi^2(1)$<1).

3. SUMMARY AND DISCUSSION

The present study examined how speech-in-speech recognition accuracy is influenced by variation in the target-background timing relationship at the local contextual level, i.e. within the time-frame of individual test trials but outside of the time-frame of the to-be-recognized sentence. Specifically, the relative timing of the onset and offset of the target and background speech was manipulated. Importantly, the approach in this study involved comparisons across conditions with controlled energetic masking characteristics within the time-frame of the to-be-recognized sentence so that the influence of local context could be isolated from the influence of energetic overlap between the target and background speech signals.

Three main findings emerged from this study. First, results replicated the mismatched language benefit [6]. That is, native English listeners showed better recognition of English target sentences when presented with background speech in a different language (i.e. Dutch) compared to when presented with background speech in the same language as the target speech (i.e. English). This release from masking, which was based on the target-background language mismatch, amounted to a benefit of approximately 38% (15 percentage points from the baseline of 61% correct recognition). Second, these data showed that synchronous onset and offset of the target and the background speech increased recognition of English-in-English sentences relative to asynchronous target and background timing. This synchrony-based release from masking amounted to a benefit of approximately 25% (15 percentage points from the baseline of 61% correct recognition). Finally, we note that there was no effect of synchronicity for English-in-Dutch recognition: for the English-in-Dutch trials, the speech recognition rate was stable at 84% correct recognition regardless of synchronous or asynchronous target and background timing.

A possible account for this pattern of results is that familiarity with the English background lead in the asynchronous English-in-English condition might have attracted attention towards the English background. That is, once a familiar language was recognized in the background speech stream, the speech recognition system may have remained attuned to that stream as a potential source of communicatively relevant information. In the English-in-English synchronous condition, however, the background stream may not have had sufficient exposure to build up a separate stream, thereby conferring a recognition benefit for the target speech stream. In contrast, lack of familiarity with the Dutch background lead in the asynchronous English-in-Dutch condition might have turned attention away from the Dutch background (now recognized as an uninformative speech stream) and towards the English target. The equivalent performance on the English-in-Dutch synchronous and asynchronous conditions may then be primarily determined by the target-background acoustic relationship within the time-frame of the to-be-recognized sentence, which is invariant across the two conditions.

Note that a critical feature of the current study's design was the constant amount of energetic masking across all of the critical comparisons. While this establishes that the influence of local context (i.e. onset and offset asynchrony versus synchrony) for English-in-English recognition is independent of energetic masking, it is possible that this feature of the study placed a limit on the range of performance variation available for experimental manipulation. For example, a drop in the signal-to-noise ratio may lower performance on the English-in-Dutch trials to a level where local contextual effects would be revealed as they were for the English-in-English trials in the present study. This would then indicate that local contextual effects are modulated by energetic masking effects with local effects being more salient under relatively high energetic masking.

In conclusion, the present work demonstrated that speech-in-speech recognition accuracy is sensitive to variation in the local context of the target-background signals. In particular, we observed a release from masking for English-in-English recognition when the target and background were played out with synchronous rather than asynchronous onsets and offsets. Future research is needed to identify the mechanism(s) that underlies this release from masking and its relationship to energetic masking.

4. REFERENCES

[1] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Memory Lang.* 59, 390-412.

[2] Barr, D. J. 2008. Analyzing 'visual world' eye tracking data using multilevel logistic regression. *J. Memory Lang.* 59, 457-474.

[3] Barr, D., Levy, R., Scheepers, C., Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68(3), 255-278.

[4] Bench, J., Kowal, A., Bamford, J. 1979. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology* 13, 108-112.

[5] Brouwer, S.M., Bradlow, A.R. (2014). Contextual variability during speech-in-speech recognition. *JASA-EL* 136(1), EL26-EL32.

[6] Brouwer S., Van Engen K.J., Calandruccio L., Bradlow A.R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. *JASA* 131(2), 1449-64.

[7] Brungart, D. S., Simpson, B. D., Ericson, M. A., Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *JASA 110*(5), 2527-2538.

[8] Freyman, R. L., Balakrishnan, U., Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *JASA* 109 (5 Pt 1), 2112-2122.

[9] Garcia Lecumberri, M. L., Cooke, M. (2006). Effect of masker type on native and nonnative consonant perception in noise. *JASA* 119(4), 2445-2454.

[10] IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. 1969. *IEEE Transactions on Audio and Electroacoustics* 17, 227-246.